# Vinum Analytica: Wine Review Insight

Paolo Palumbo

## Abstract

This paper presents a data mining project that analyzes the language used in wine reviews written by sommeliers. We compare the performance of three different models: Decision Tree, Random Forest, and Neural Network. The analysis includes a detailed examination of preprocessing techniques, class contamination removal, and feature vectorization using TF-IDF. We assess the models based on their accuracy and present the results through various evaluations.

## 1 Introduction

The growing availability of online wine reviews provides a valuable resource for understanding consumer preferences and trends. This study aims to analyze the language used in these reviews to derive insights that can help in predicting wine ratings. We employ three machine learning models—Decision Tree, Random Forest, and Neural Network—to classify and predict wine ratings based on text features extracted from reviews.

## 2 Preprocessing

Data preprocessing is a critical step in any machine learning pipeline. In this project, we applied several preprocessing techniques to prepare the text data for modeling:

- **Text Normalization:** All text was converted to lowercase to ensure uniformity.
- **Tokenization:** Text was split into individual words or tokens.
- **Stop Words Removal:** Common words that do not contribute to the semantic meaning were removed.
- **Stemming/Lemmatization:** Words were reduced to their base or root form.

## 3 Feature Vectorization

To transform the text data into a numerical format suitable for machine learning, we used the TF-IDF (Term Frequency-Inverse Document Frequency) method. This approach captures the importance of words in the context of the entire dataset and helps in feature extraction.

## 4 Modeling and Hyperparameter Tuning

We divided the dataset into training and test sets. The training set underwent oversampling using SMOTE to address class imbalance, with a limit of 200,000 records. We conducted a hyperparameter search using 6-fold cross-validation for each model to determine the best parameters based on accuracy.

### 4.1 Decision Tree

### 4.2 Random Forest

### 4.3 Neural Network

## 5 Results

We evaluated the performance of the models using accuracy as the primary metric. The results for each

| Index | Criterion | Min Impurity Decrease | M |
|---|---|---|---|
| 0 | log_loss | 0.0 | |
| 1 | entropy | 1e-12 | |
| 2 | gini | 0.0001 | |
| 3 | entropy | 0.0001 | |
| 4 | entropy | 1e-12 | |
| 5 | log_loss | 0.0001 | |
| 6 | gini | 1e-08 | |
| 7 | entropy | 1e-08 | |

Hyperparameters for Decision Tree

| Index | n_estimators | criterion | max_depth |
|---|---|---|---|
| 0 | 150 | log_loss | None |
| 1 | 50 | gini | None |
| 2 | 50 | gini | 1000 |
| 3 | 100 | entropy | None |
| 4 | 50 | entropy | None |
| 5 | 100 | gini | 1000 |
| 6 | 50 | log_loss | None |
| 7 | 100 | gini | None |

Hyperparameters for Random Forest

| Index | Hidden Size | Epochs | Learning Rate |
|---|---|---|---|
| 0 | 64 | 10 | 0.005 |
| 1 | 64 | 5 | 0.001 |
| 2 | 64 | 15 | 0.005 |
| 3 | 16 | 5 | 0.01 |
| 4 | 16 | 5 | 0.005 |
| 5 | 64 | 5 | 0.005 |
| 6 | 16 | 15 | 0.01 |
| 7 | 64 | 10 | 0.01 |

Hyperparameters for Neural Network

model are presented through violin plots, which illustrate the distribution of accuracy scores across different hyperparameters.

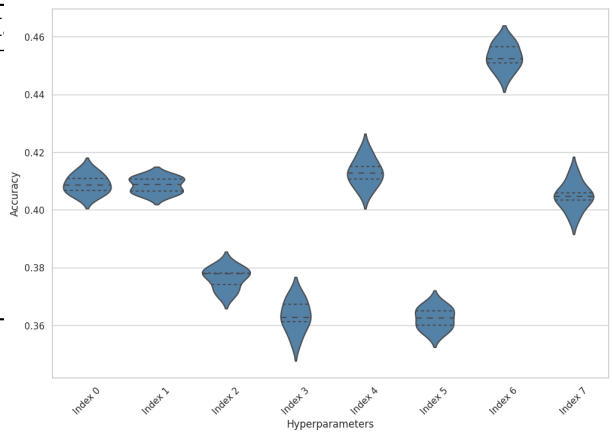


Figure 1: Violin Plot for Decision Tree

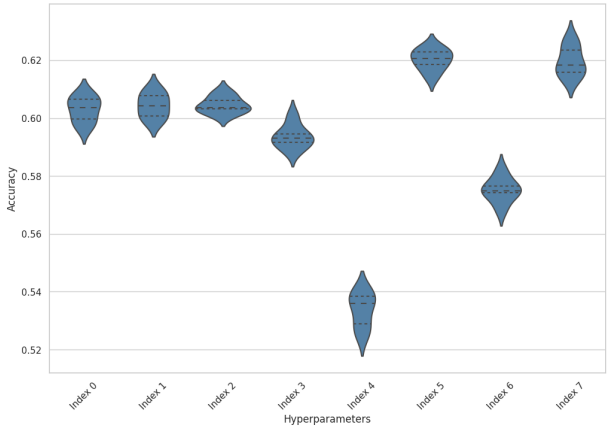


Figure 2: Violin Plot for Random Forest

# 6 Discussion

The analysis reveals the comparative performance of the Decision Tree, Random Forest, and Neural Network models in predicting wine ratings based on text reviews. The violin plots provide insights into the stability and variability of each model's accuracy across different hyperparameter settings. The Random Forest model generally showed the highest accuracy, followed by the Neural Network and Decision Tree models.

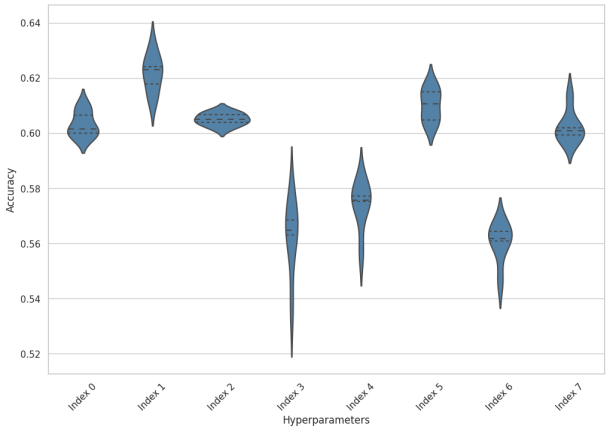


Figure 3: Violin Plot for Neural Network

# 7 Conclusion

In this study, we have successfully compared three machine learning models for predicting wine ratings using text data. The Random Forest model demonstrated superior performance, while the Neural Network and Decision Tree models offered valuable insights as well. Future work could explore additional models and advanced preprocessing techniques to further improve prediction accuracy.